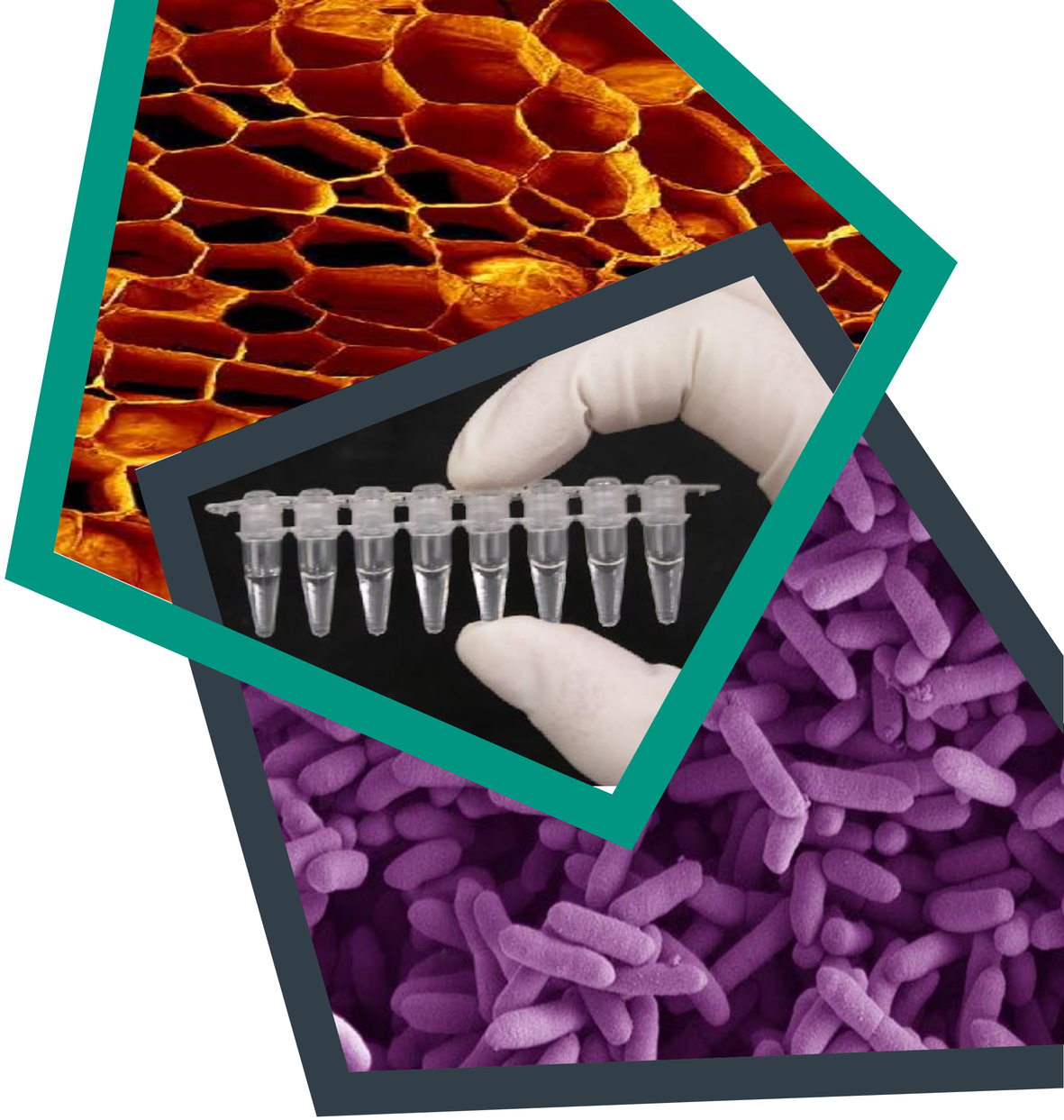
Quadram Institute
Science Health
Food Innovation

A primer on metabarcoding

Andrea Telatin, Gut Microbial Health
Quadram Institute, UK

Quadram Institute
Science Health
Food Innovation

What is metabarcoding about?



Metabarcoding

"Gene sequences,
most commonly those encoding rRNAs,
provide a basis for
estimating microbial phylogenetic diversity
and generating taxonomic inventories of
[...] microbial populations."

– Sogin *et. al* 2006

Questions

- Who is in the sample?
- What are they doing?

Questions

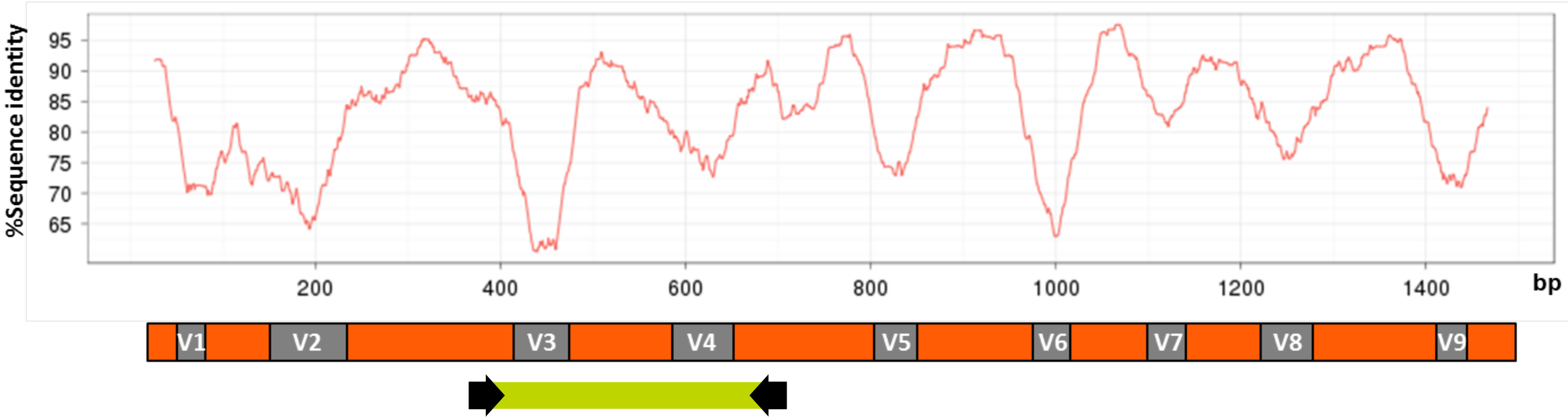
- Who is in the sample?

  16S/ITS Amplicons, Metagenomics

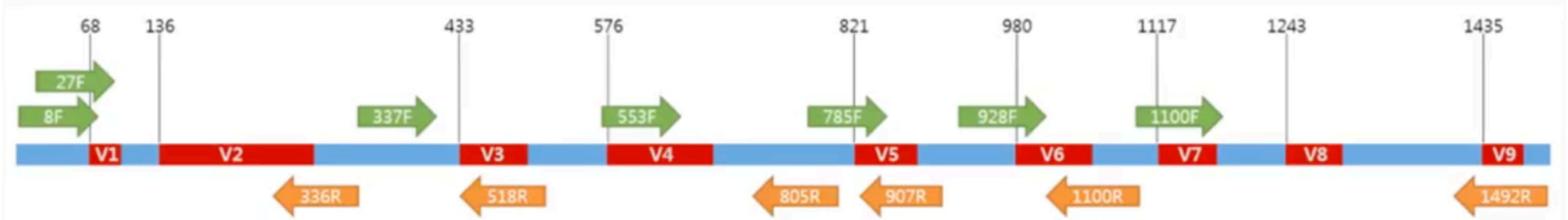
- What are they doing?

  Metagenomics

Anatomy of a gene: 16S rDNA



Anatomy of a gene: 16S rDNA



V1-V3 ~510 bp for Roche 454

V3-V4 ~428 bp for MiSeq PE

V3-V5 ~548 bp for Roche 454

V4 ~252 bp for HiSeq

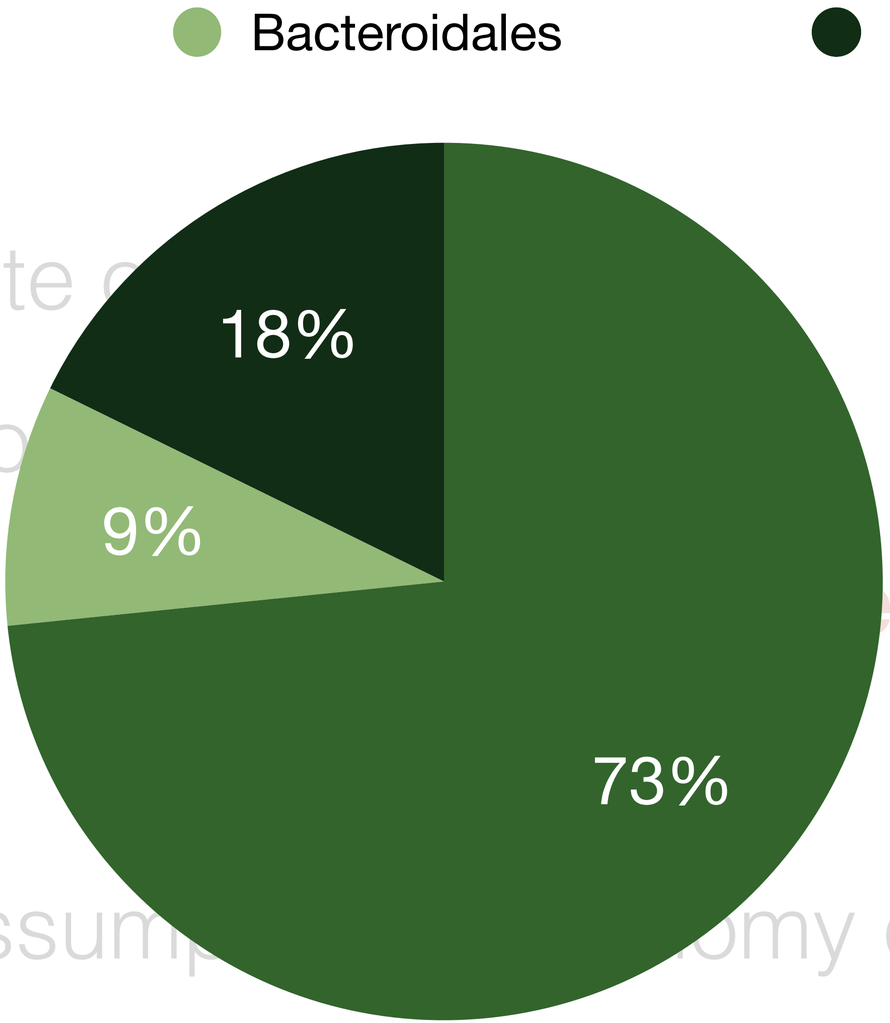
~562 bp for Roche 454 V6-V9

V1-V9 (Full-length)

Biases?

- Sampling (replicate or lie)
- Cell breaking (spores, cell wall,…)
- Amplify (primers? cycles? chimeric molecules)
- Sequencing
- Bioinformatics (assumptions, taxonomy databases…)

Biases?



- Sampling (replicate c
- Cell breaking (spo
- Amplify (primers? ecules)
- Sequencing
- Bioinformatics (assump omy databases…)

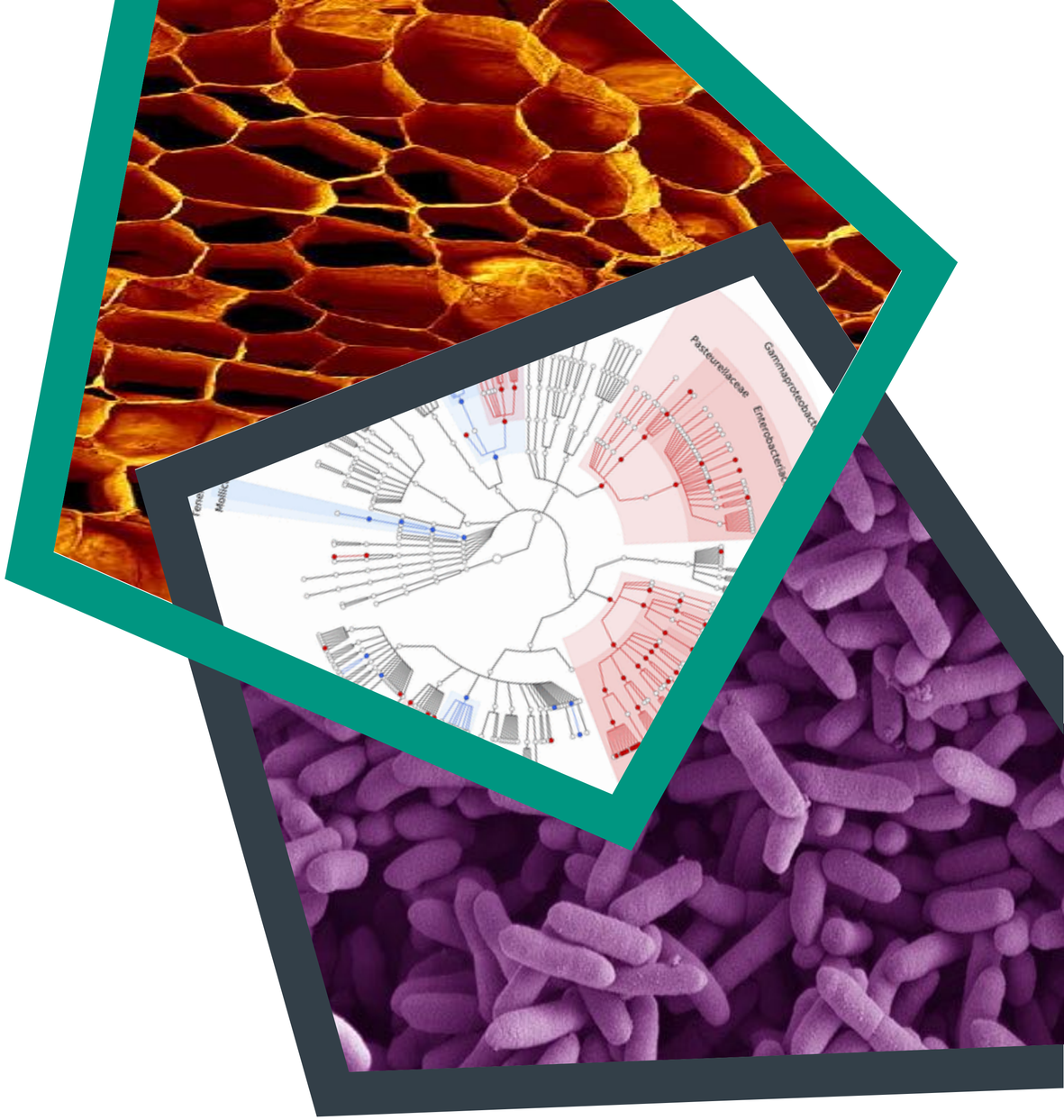
Biases?

That's why we **compare** different samples

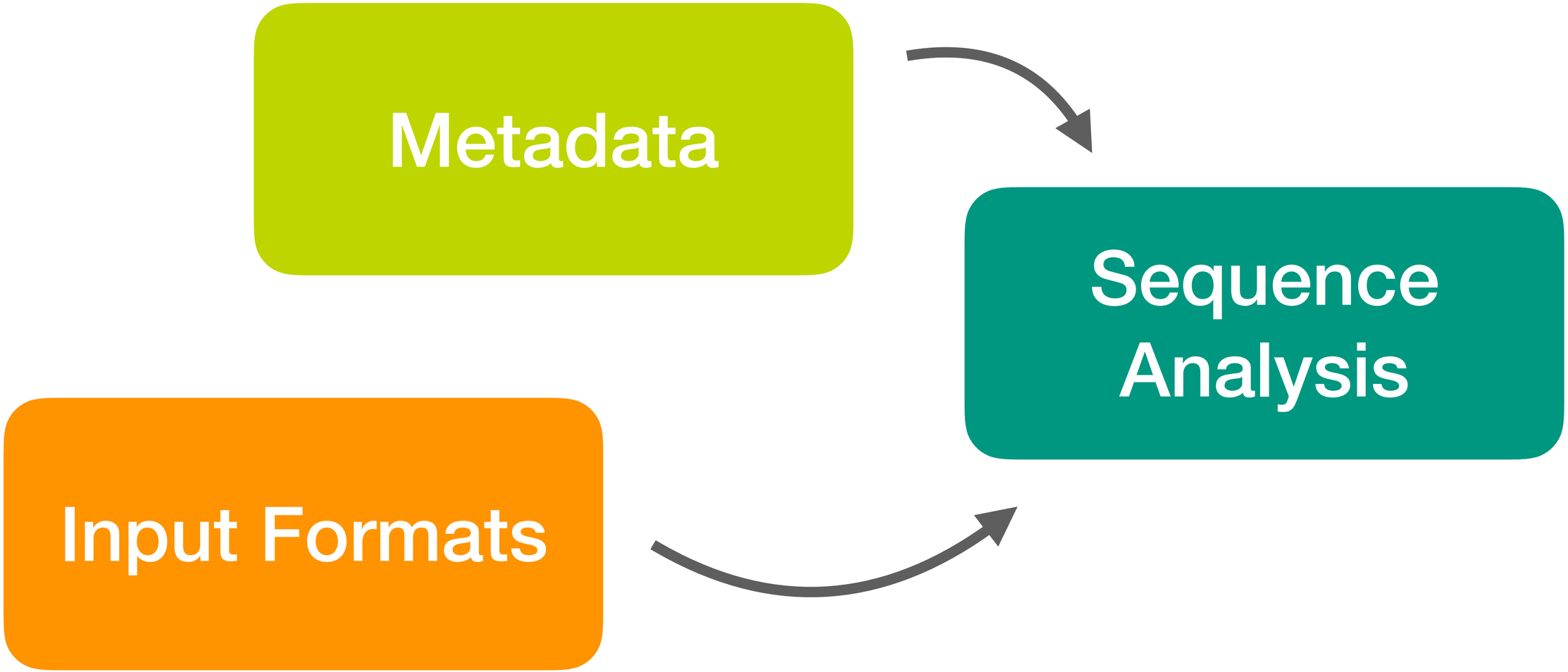
- Sampling (replicate or lie)
- Cell breaking (spores, cell wall,…)
- Amplify (primers? cycles? chimeric molecules)
- Sequencing
- Bioinformatics (assumptions, taxonomy databases…)

Quadram Institute
Science Health
Food Innovation

Bioinformatics analysis



Tasks



How (tools)

QIIME 1.9

LOTUS

USEARCH

QIIME 2

MOTHUR

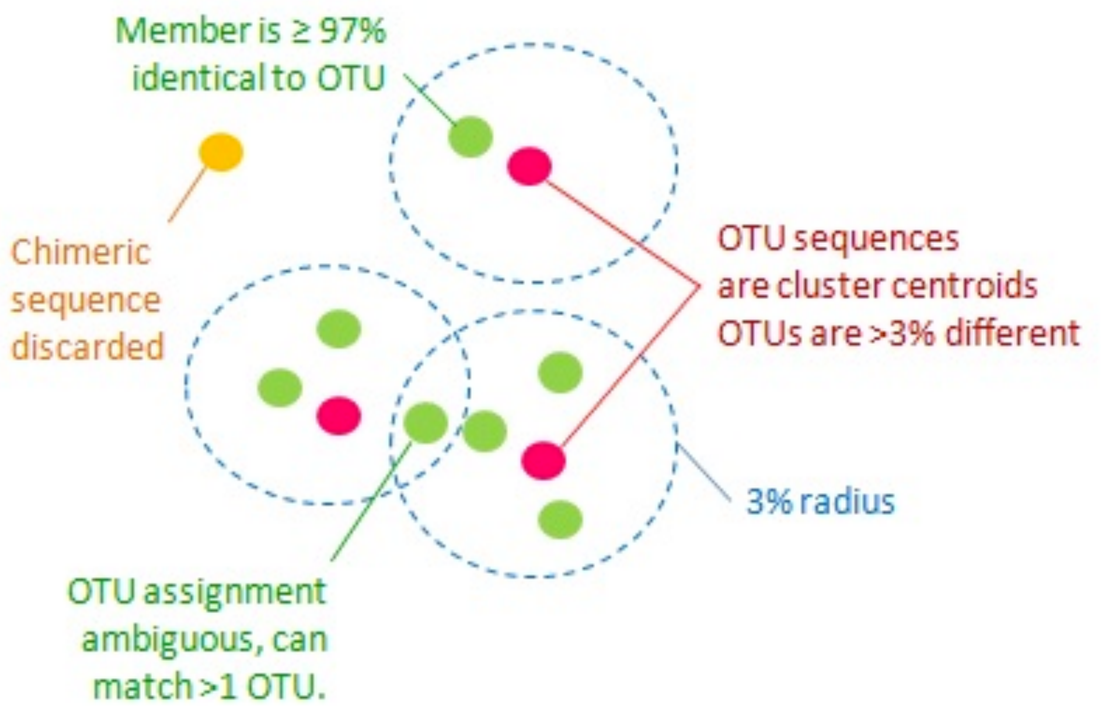
...

A training source: USEARCH

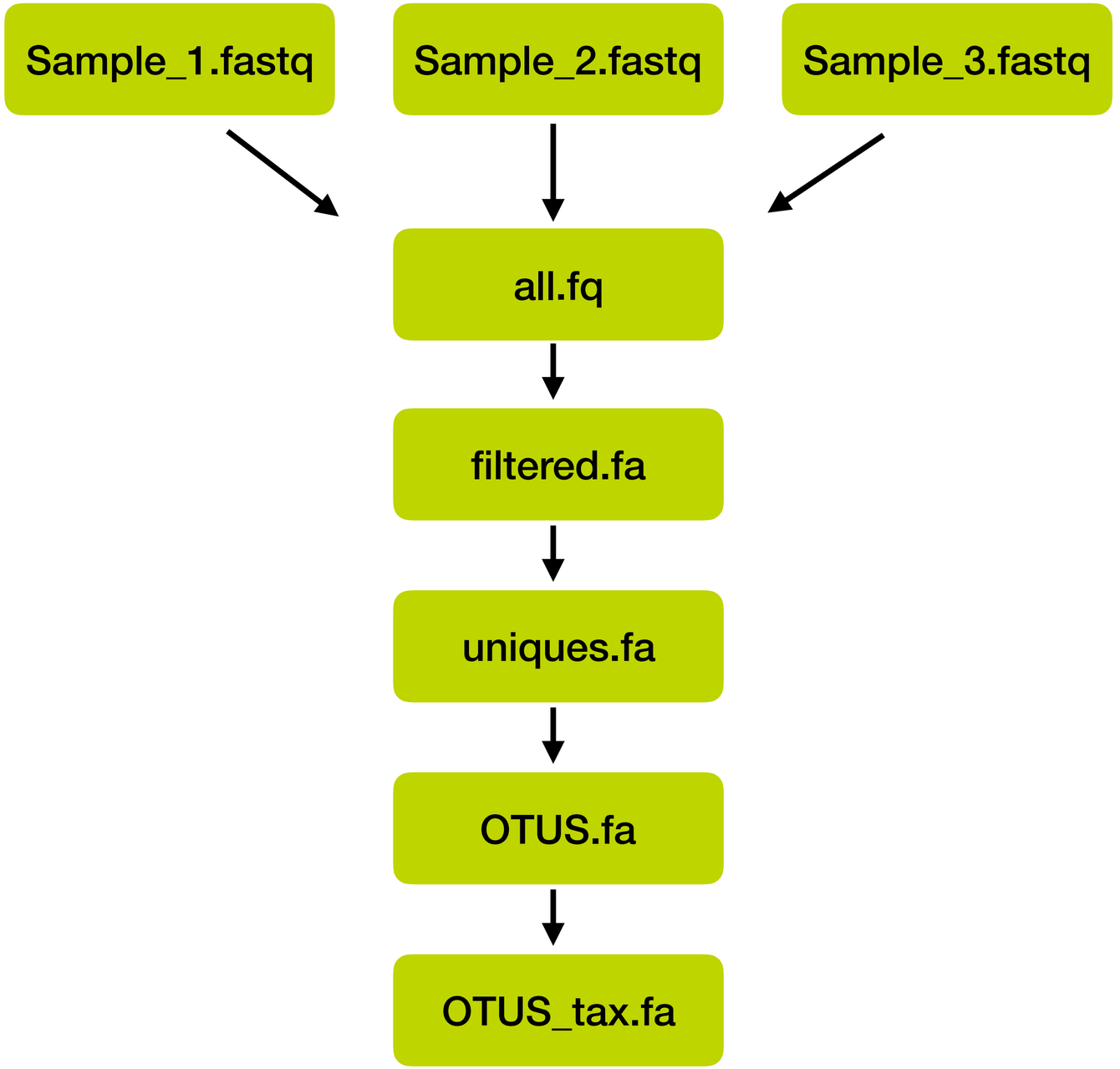
USEARCH is a monolithic package written by Robert Edgar, an independent bioinformatics researcher.

Can be used to see how we manipulate reads to make an annotated OTU table.

32-bit free download: **http://drive5.com/**

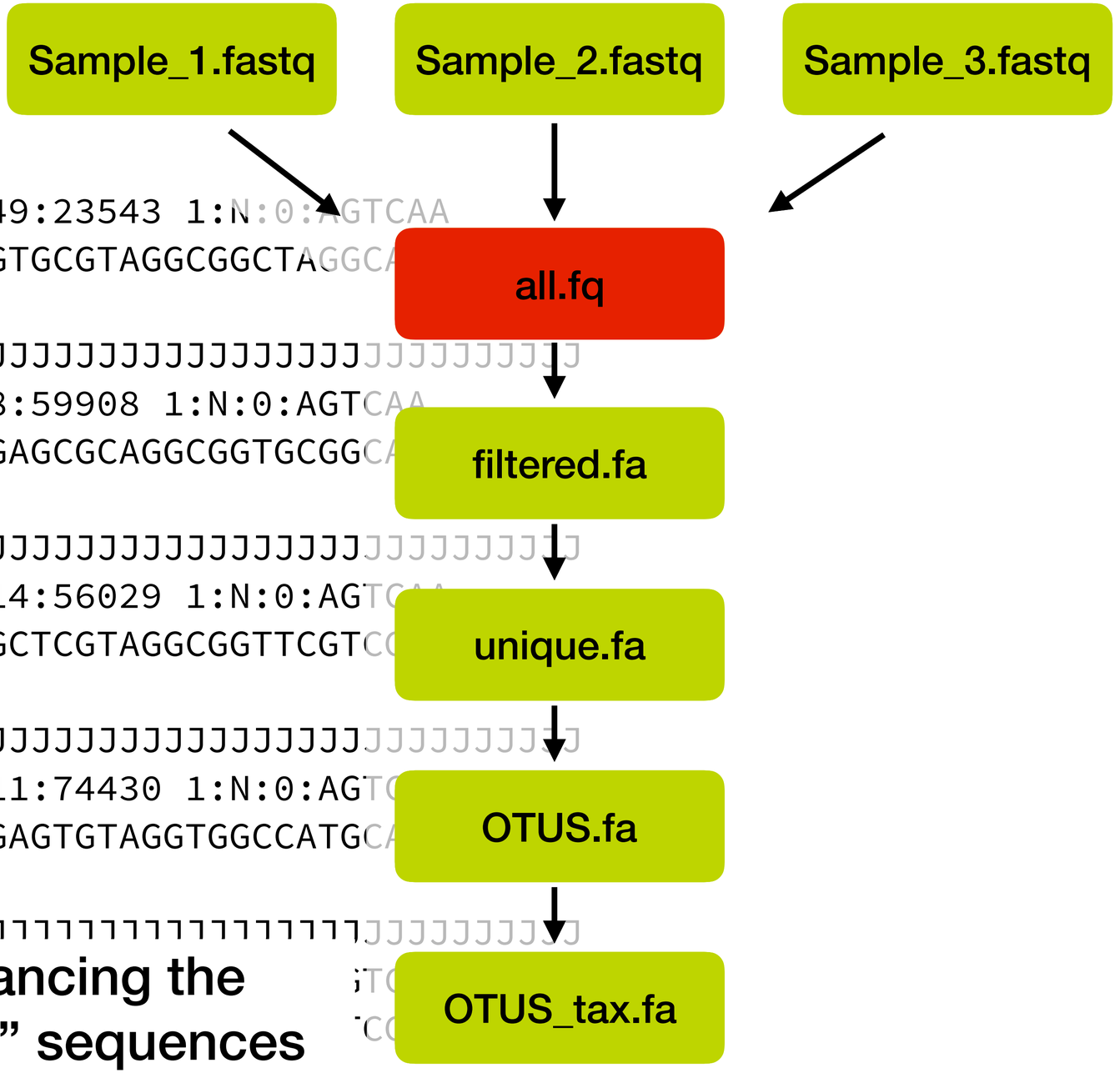


A simplified example



http://drive5.com/usearch/manual/pipe_examples.html

Combine reads



tracing sample IDs

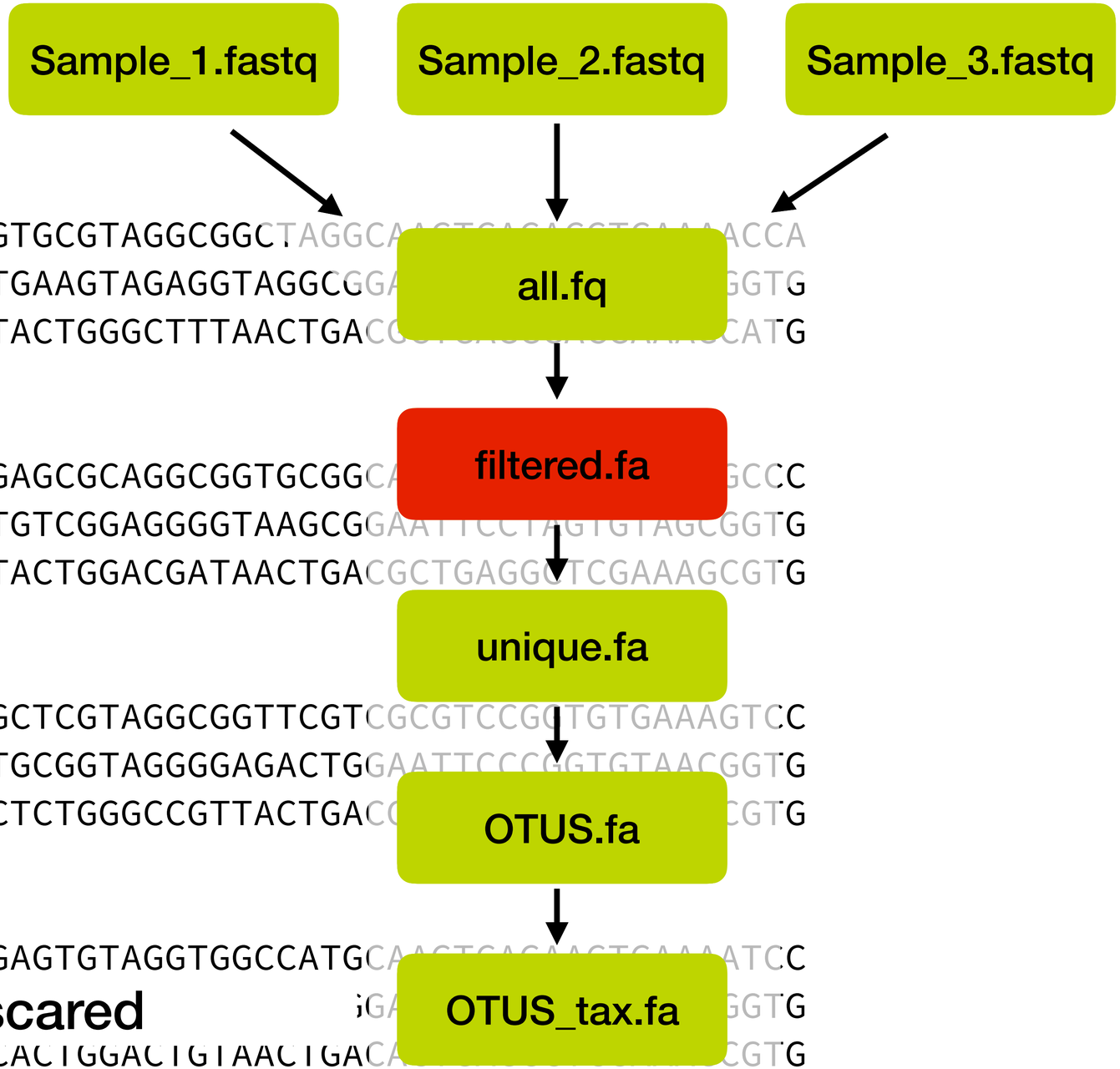
```
@1.control::D00420:186:HF5VMBCX2:1:1103:13149:23543 1:N:0:AGTCAA
TACGTAGGGAGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGTGCGTAGGCGGCTAGGCA
+
IIIAG?GGIIIGIIGIGGGGGJJJJJJJJJJJJJJJJJJ?JJJJJJJJJJJJJJJJJJJJJJ
@1.control::D00420:186:HF5VMBCX2:1:2203:5458:59908 1:N:0:AGTCAA
TACGTAGGGAGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGCA
+
GIIIIGGAGGGIIIIIIGI<GGJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@1.control::D00420:186:HF5VMBCX2:1:2113:15314:56029 1:N:0:AGTCAA
TACGTAGGGTGCAAGCGTTATCCGGAATTATTGGGCGTAAAGGGCTCGTAGGCGGTTCGTCG
+
GGIIGIIIIIGIIIIGIIGIIGJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@1.control::D00420:186:HF5VMBCX2:1:2203:17011:74430 1:N:0:AGTCG
AACGTATGGAGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGTGTAGGTGGCCATGCA
+
IIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIJJJJJJJJJJ
```

this is not necessary, but help enhancing the signal/noise ration in detecting “rare” sequences

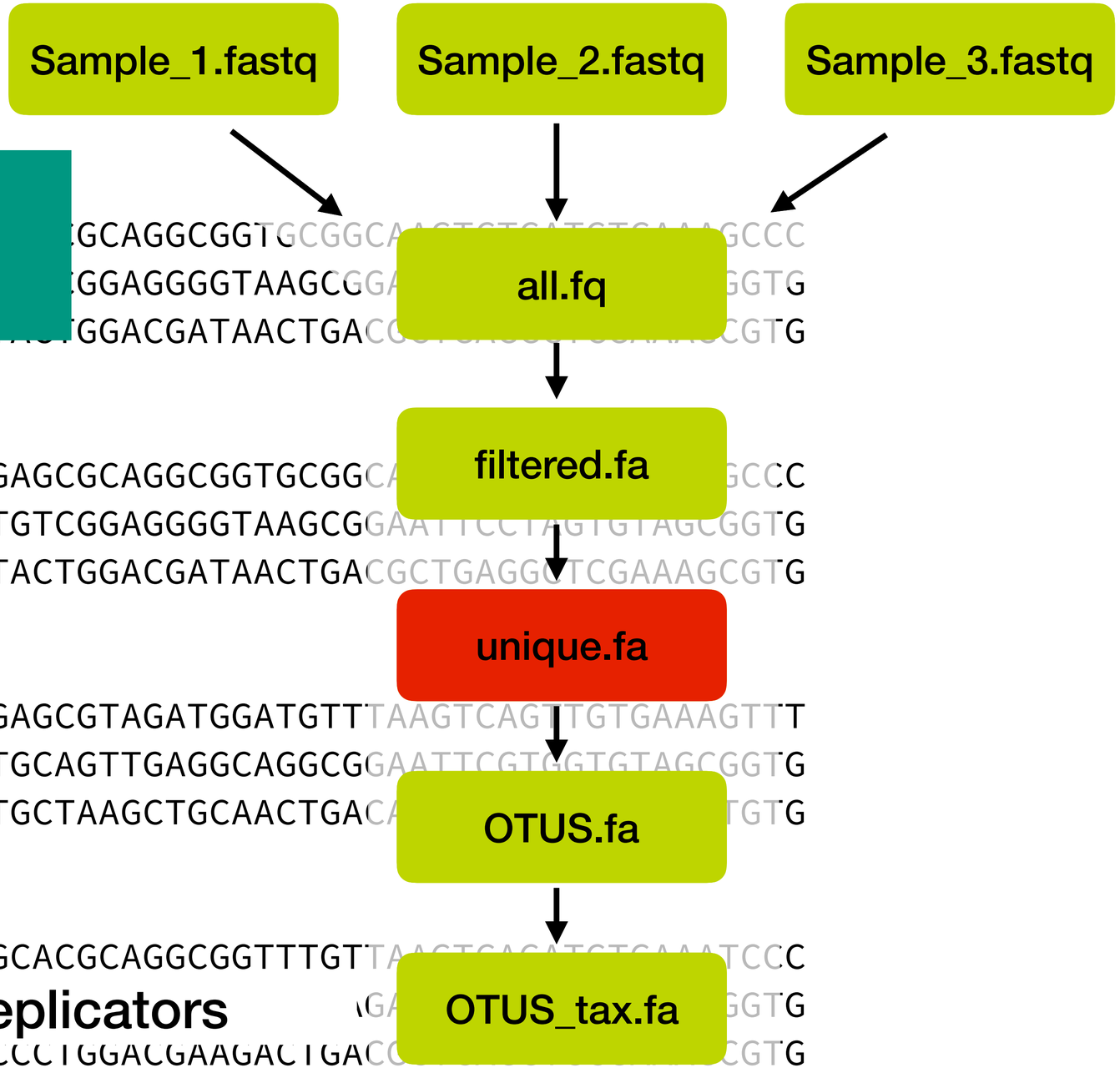
Quality filter

```
>Filt1
TACGTAGGGAGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGTGCGTAGGCGGCTAGGCAAGTCAGAGGTGAAAACCA
TGGGCTCAACCTGTGGACTGCGGTTGAAACTGTTTAGCTTGAGTGAAGTAGAGGTAGGCGGAAATTCCTAGTGTAGCGGTG
AAATGCGTAGAGATCGGGAGGAACACCAGTGGCGAAGGCGGCCTACTGGGCTTTAACTGACGCTGAGGCTCGAAAGCATG
GGTAGCAAACAGGATTAGAAACCCGCGTAGTCC
>Filt2
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGCAAGTCAGAGGTGAAAGCCC
GGGGCTCAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTCGGAGGGGTAAGCGGAATTCCTAGTGTAGCGGTG
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGGCTTACTGGACGATAACTGACGCTGAGGCTCGAAAGCGTG
GGGAGCAAACAGGATTAGAAACCCTGGTAGTCC
>Filt3
TACGTAGGGTGCAAGCGTTATCCGGAATTATTGGGCGTAAAGGGCTCGTAGGCGGTTCGTCGCGTCCGCTGTGAAAGTCC
ATCGCTTAACGGTGGATCTGCGCCGGGTACGGGCGGGCTGGAGTGCGGTAGGGGAGACTGGAATTCCCGGTGTAACGGTG
GAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGCAGGTCTCTGGGCCGTTACTGACGCTGAGGCTCGTAAAGCGTG
GGGAGCGAACAGGATTAGAAACCCCCGTAGTCC
>Filt4
AACGTATGGAGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGTGTAGGTGGCCATGCAAGTCAGAAGTGAAAATCC
AGAATGTGGTAGATATCGGGAAGAACACCAGTGGCGAAGGCGGCTCACTGGACTGTAACTGACACTGAGGCTCGTAAAGCGTG
```

note that quality values are discared



Dereplication

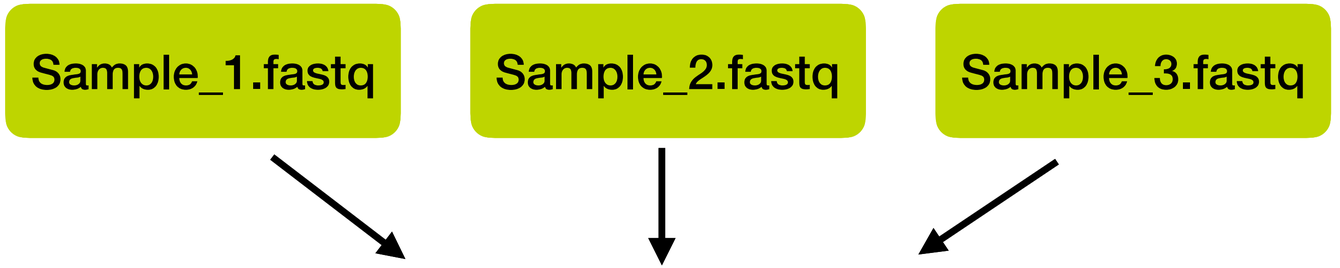


propagating "metadata"

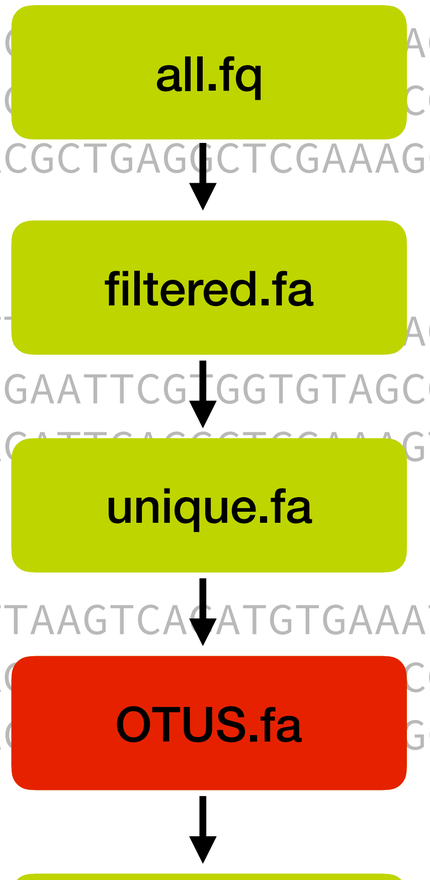
```
>Uniq1;size=6103;
TACGTATGGTGCAAGCGTTA
GGGGCTCAACCCCGGTACTGCA
AAATGCGTAGATATTAGGAGGA
GGGAGCAAACAGGATTAGAAACCCCCGTAGTCC
>Uniq2;size=5788;
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGC
GGGGCTCAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTCGGAGGGGTAAGCG
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGGCTTACTGGACGATAACTGA
GGGAGCAAACAGGATTAGATACCCCCGTAGTCC
>Uniq3;size=5082;
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTT
GCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCG
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGA
GGTATCAAACAGGATTAGAAACCCCCGTAGTCC
>Uniq4;size=3438;
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGT
```

"sdm" and "usearch" are fast dereplicators

OTU/ASV picking

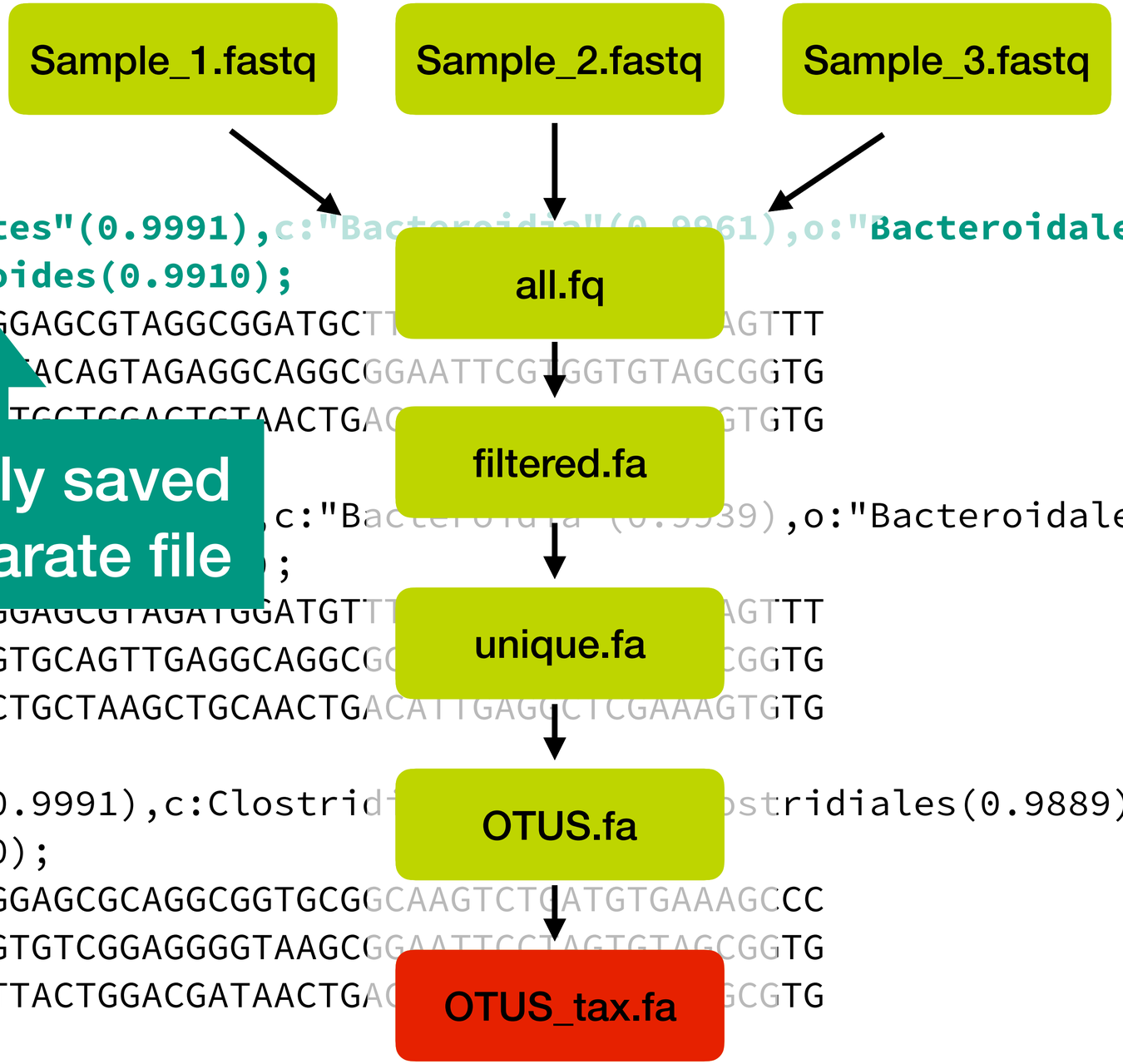



```
>Otu1
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGG                AGCCC
GGGGCTCAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTCGGAGGGGTAAGCGG                CGGTG
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGGCTTACTGGACGATAACTGACGCTGAGGCTCGAAAGCGTG
GGGAGCAAACAGGATTAGAAACCCCCGTAGTCC
>Otu2
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTT                AGTTT
GCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTG
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTGCAAAGTGTG
GGTATCAAACAGGATTAGAAACCCCCGTAGTCC
>Otu3
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCC
CGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAG                CGGTG
                                                                              GCGTG
                                                                              AGTCC
ATCGCTTAACGGTGGATCTGCGCCGGGTACGGGCGGGCTGGAGTGCGGTAGGGGAGACTG                CGGTG
GAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGCAGGTCTCTGGGCCGTTACTGACGCTGAGGAGCGAAAGCGTG
```



We can cluster sequences using identity (OTU), or use sequencer-specific denoisers to detect high quality sequence representatives (ASV)

Classification



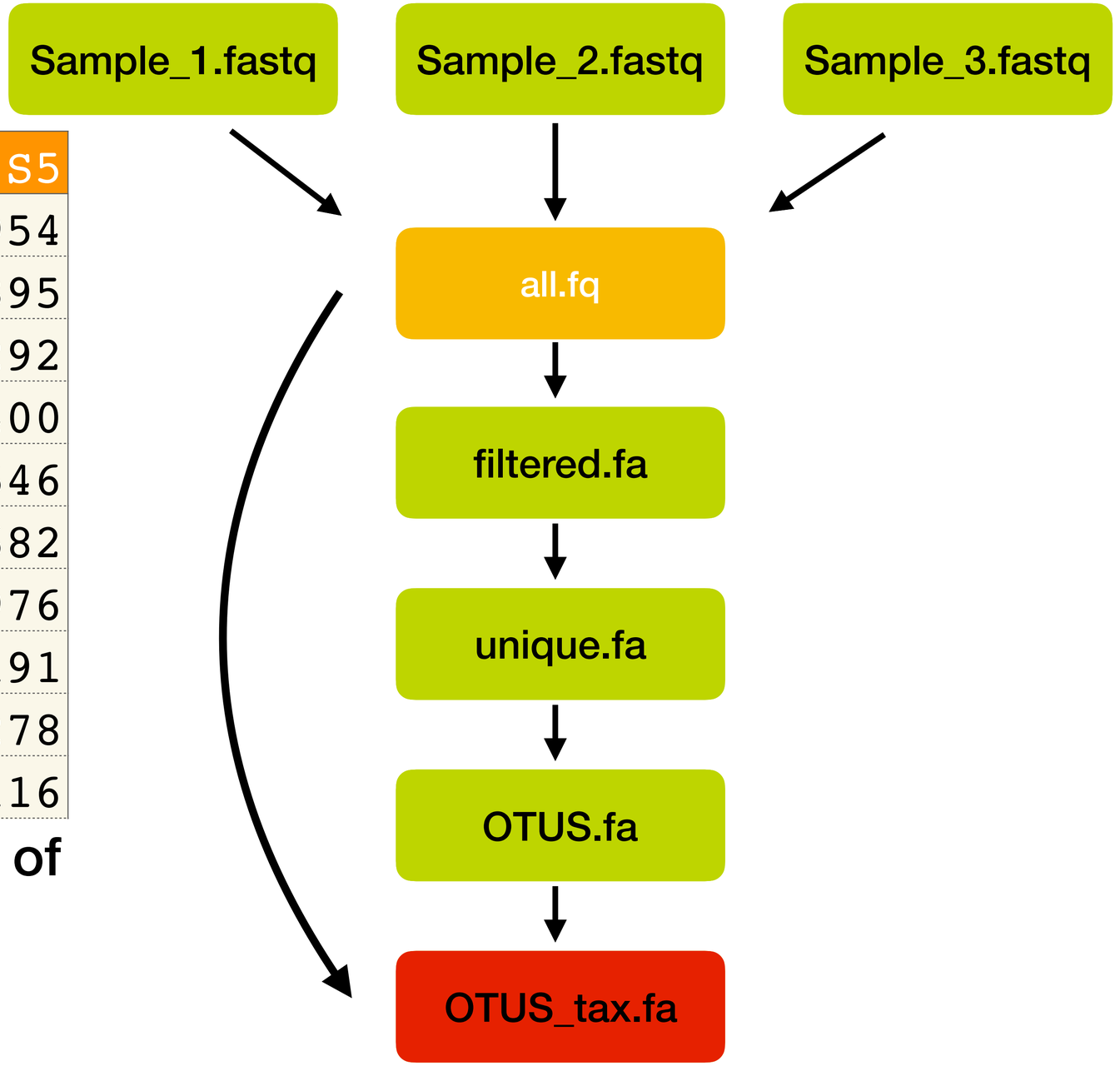
commonly saved as a separate file

```
>Otu8;tax=d:Bacteria(1.0000),p:"Bacteroidetes"(0.9991),c:"Bacteroidia"(0.9961),o:"Bacteroidale
(0.9952),f:Bacteroidaceae(0.9910),g:Bacteroides(0.9910);
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGATGCTT                AGTTT
GCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTGTCTTG  ACAGTAGAGGCAGGCGGAATTCGTGGTGTAGCGGTG
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGC TGCTGGACTGTAACTGAC               GTGTG
GGTATCAAACAGGATTAGAAACCCCCGTAG
>Otu2;tax=d:Bacteria(1.0000),p                                 c:"Bacteroidia"(0.9939),o:"Bacteroidale
(0.9939),f:Bacteroidaceae(0.99                                 );
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTT               AGTTT
GCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGG               CGGTG
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTG
GGTATCAAACAGGATTAGAAACCCCCGTAGTCC
>Otu1;tax=d:Bacteria(1.0000),p:Firmicutes(0.9991),c:Clostrid                ostridiales(0.9889)
:Lachnospiraceae(0.9573),g:Roseburia(0.5720);
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGCAAGTCTGATGTGAAAGCCC
GGGGCTCAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTCGGAGGGGTAAGCGGAATTCCTAGTGTAGCGGTG
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGGCTTACTGGACGATAACTGAC               GCGTG
GGGAGCAAACAGGATTAGAAACCCCCGTAGTCC
```

The OTU table

#OTU	S1	S2	S3	S4	S5
Otu1	14303	347	17513	474	14954
Otu2	10636	1312	7312	1486	11395
Otu4	3229	2662	7050	3612	3292
Otu3	382	9054	467	10925	400
Otu5	6523	1001	8289	1446	6546
Otu9	679	5963	1058	8398	682
Otu23	5025	413	6982	830	4976
Otu6	1213	6380	570	10908	1191
Otu7	2172	922	3836	1202	2278
Otu11	46	2386	70	2458	116

Produced counting how many reads of each sample map unambiguously against a specific OTU/ASV



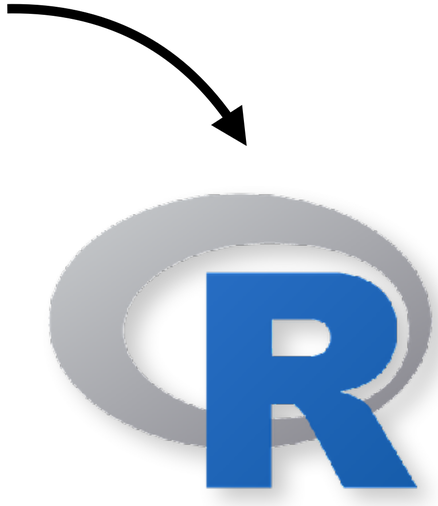
From the OTU table to numeRical ecology

#OTU	S1	S2	S3	S4	S5	taxonomy
Otu1	14303	347	17513	474	14954	d:Bacteria;p:Firmicutes
Otu2	10636	1312	7312	1486	11395	d:Bacteria;p:Bacteroidetes;o:Bactero…
Otu4	3229	2662	7050	3612	3292	d:Bacteria;p:Proteobacteria;o:Alpha…
Otu3	382	9054	467	10925	400	d:Bacteria;p:Bacteroidetes
Otu5	6523	1001	8289	1446	6546	d:Bacteria;p:Firmicutes;o:Clostridia
Otu9	679	5963	1058	8398	682	d:Bacteria;p:Bacteroidetes;…
Otu23	5025	413	6982	83		
Otu6	1213	6380	570	1090		
Otu7	2172	922	3836	120		
Otu11	46	2386	70	245		
Otu8	3911	2136	3706	1997	4082	d:Bacteria;p:Proteobacteria;o:Gamma…

OTU table + Taxonomy

	Diet	Sex	Date
S1	Normal	M	21-04
S2	Normal	F	22-04
S3	LowFat	F	21-04
S4	Lo		5
S5	Bi		4

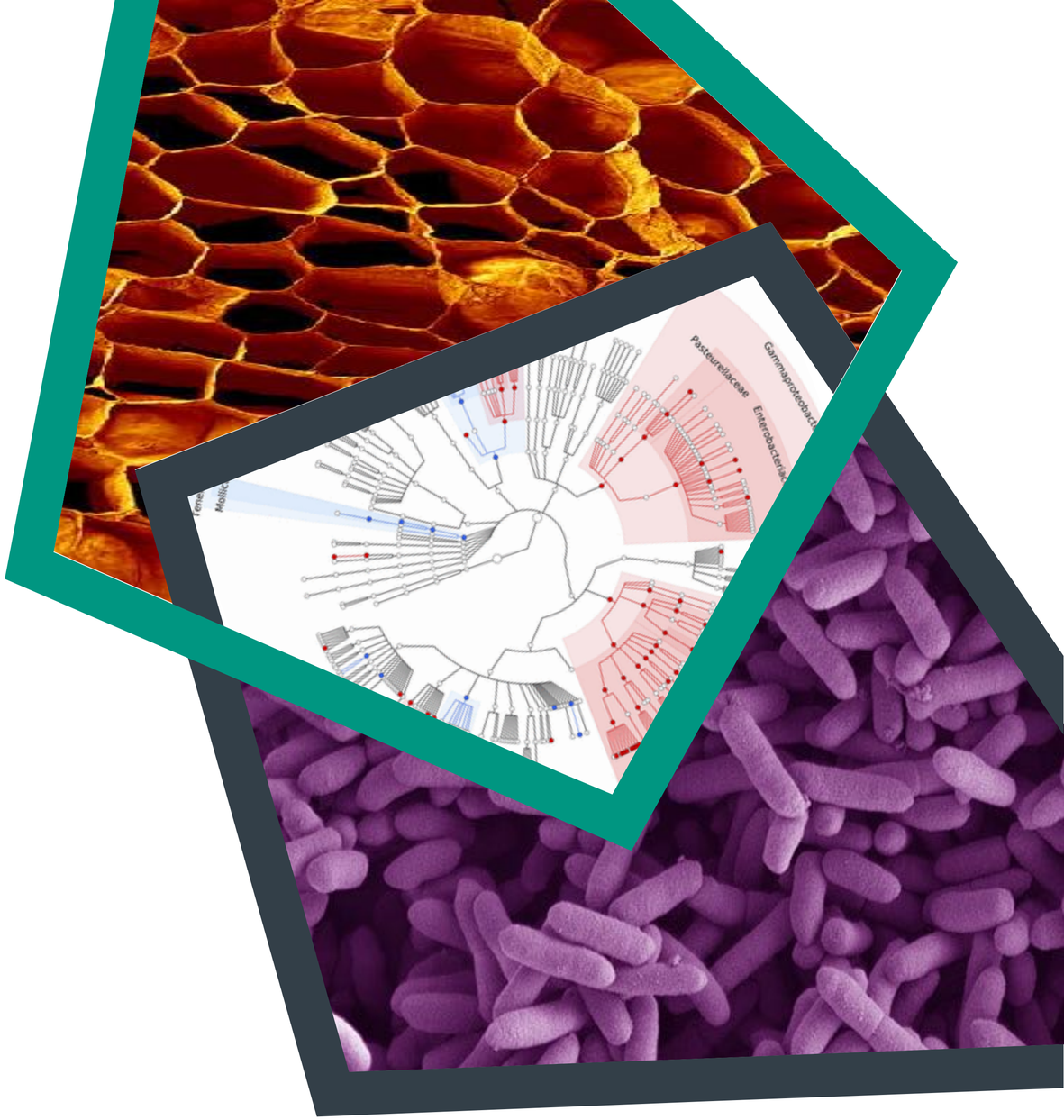
Metadata Table



- Normalization
- Alpha Diversity
- Beta Diversity
- Differential OTU ab.
- …

Quadram Institute
Science Health
Food Innovation

Packages for metabarcoding analysis



Qiime 1.9

… an open-source bioinformatics pipeline to perform microbiome analysis from raw DNA sequencing data, now discontinued

- Rob Knight group and collaborators
- A collection of wrappers around well known tools (like UPARSE)
- The wrapping is **consistent** and makes **easier** to perform the whole analysis
- From sequences to **charts**
- It has been a very popular package but developed stopped years ago to produce a complete rewrite (Qiime 2)

www.qiime.org

Qiime 1.9

- With default parameters and database, was commonly overestimating diversity (i.e. several spurious OTUs identified)
- Now discontinued and should be avoided
- Still used for “compatibility” with legacy analyses

Metadata file (Mapping File in Qiime 1 jargon)

A TSV file linking Samples with barcodes (for demultiplexing) and other user defined columns

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Description
1	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
2	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
3	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
4	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
5	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
6	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
7	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
8	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
9	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
10	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
11	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
12	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
13	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
14	NNNNNNNNNNNN	CCCCGTCAATTCMTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased

Qiime 2

- a complete rewrite of Qiime 1; introduce the new concept of **artifacts** (input/output packages with metadata)
- strongly advocates **DADA2** algorighm for **ASV** detection rather than OTU picking
- new release are produced frequently, unfortunately sometimes changing some commands (not a sign of a mature and stable package)
- Still producing charts, in my opinion sometimes worse than in the past
- Robust **traceability** and reproducibility at the expense of ease of use and interchange of data with other tools

www.qiime2.org

Qiime 2

- a complete rewrite of Qiime 1; introduce the new ca
  outp
- str
- new
  com
- Still
- Robu
  interc

```
prokka
qiime1
qiime2-2018.4
qiime2-2018.6
qiime2-2018.8
qiime2-2019.1
qiime2-2019.7
r
unicycler
virsorter
```

```
/home/ubuntu/miniconda3/envs/prokka
/home/ubuntu/miniconda3/envs/qiime1
/home/ubuntu/miniconda3/envs/qiime2-2018.4
/home/ubuntu/miniconda3/envs/qiime2-2018.6
/home/ubuntu/miniconda3/envs/qiime2-2018.8
/home/ubuntu/miniconda3/envs/qiime2-2019.1
/home/ubuntu/miniconda3/envs/qiime2-2019.7
/home/ubuntu/miniconda3/envs/r
/home/ubuntu/miniconda3/envs/unicycler
/home/ubuntu/miniconda3/envs/virsorter
```

www.qiime2.org

Qiime2 artifacts

qza

Qiime2 archive

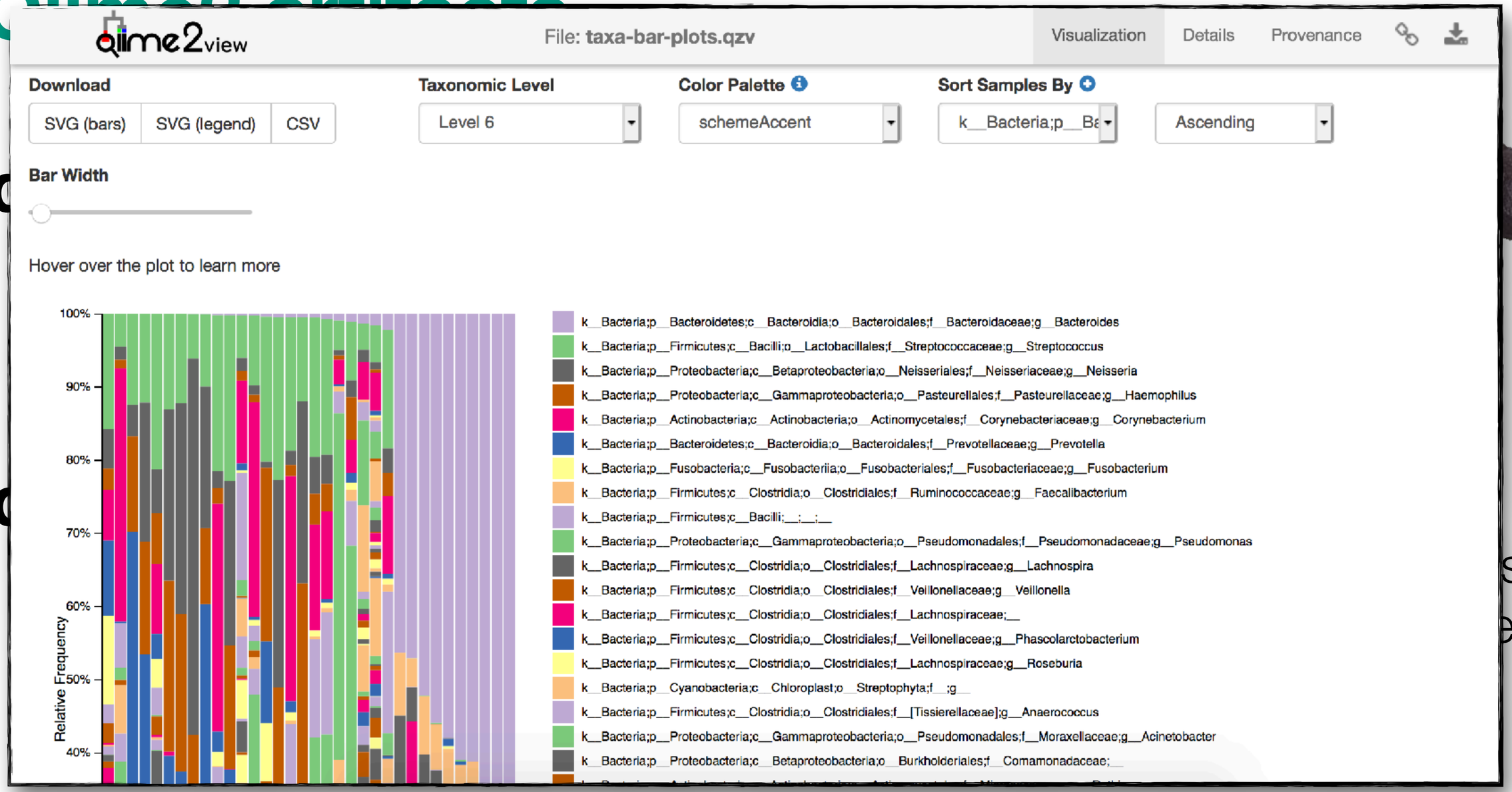
It's the output format of all Qiime2 programs. It's a ZIP files with both data and metadata.

qzv

Qiime2 visualization

It's the output format for plots/charts and tables that the user could desire to inspect. It's an HTML document (web page) embedded in a "ZIP with metadata", like qza.
http://view.qiime2.org

Qiime2 artifacts



Qiime2 command: first example

Sample01_S10_L001_R1_001.fastq
Sample01_S10_L001_R2_001.fastq
Sample02_S11_L001_R1_001.fastq
Sample02_S11_L001_R2_001.fastq

```
qiime tools import   --type 'SampleData[PairedEndSequencesWithQuality]'   \
        --input-path ./reads/   \
        --source-format CasavaOneEightSingleLanePerSampleDirFmt   \
        --output-path ./imported_reads.qza
```

imported_reads.qza

Qiime2 command: first example

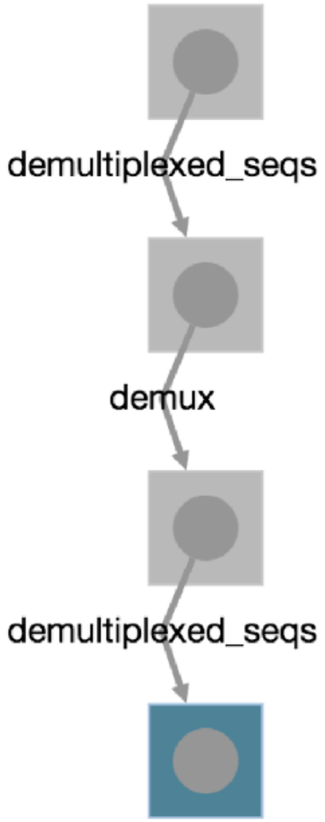
imported_reads.qza

```
$ unzip -t demux-paired-end.qza

Archive:  demux-paired-end.qza
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/VERSION   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/metadata.yaml   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/VERSION   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/metadata.yaml   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/citations.bib   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/action/action.yaml   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/64_S64_L001_R2_001.fastq.gz   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/34_S34_L001_R1_001.fastq.gz   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/60_S60_L001_R2_001.fastq.gz   OK
    testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/19_S19_L001_R1_001.fastq.gz   OK
```

“Provenance” metadata

Provenance Graph



Action Details

▼ execution:
  uuid: "4ccc8855-337b-4db6-9715-c3dd59f86251"
  ▼ runtime:
    start: 2018-03-19T08:59:26.488Z
    end: 2018-03-19T15:22:47.725Z
    duration: "6 hours, 23 minutes, 21 seconds, and 236906 microseconds"
▼ action:
  type: "method"
  plugin: "environment:plugins:deblur"
  action: "denoise_16S"
  ▼ inputs:
    ▼ 0:
      demultiplexed_seqs: "2c0cf238-bbe9-4f0c-81d3-201144a96e03"
  ▼ parameters:
    ▼ 0:
      trim_length: 450

Trying yourself

- There is a step-by-step tutorial in the Qiime2 website (below)
- At each step try “unzipping” the artifact to see that it contains standard files

https://docs.qiime2.org/2019.7/tutorials/moving-pictures/

Mothur

- a single package mainly developed by Pat Schloss
- has by default an “R-like” interface
- No charts produced, but some basic numerical ecology
- Can be slow with large datasets

www.mothur.org

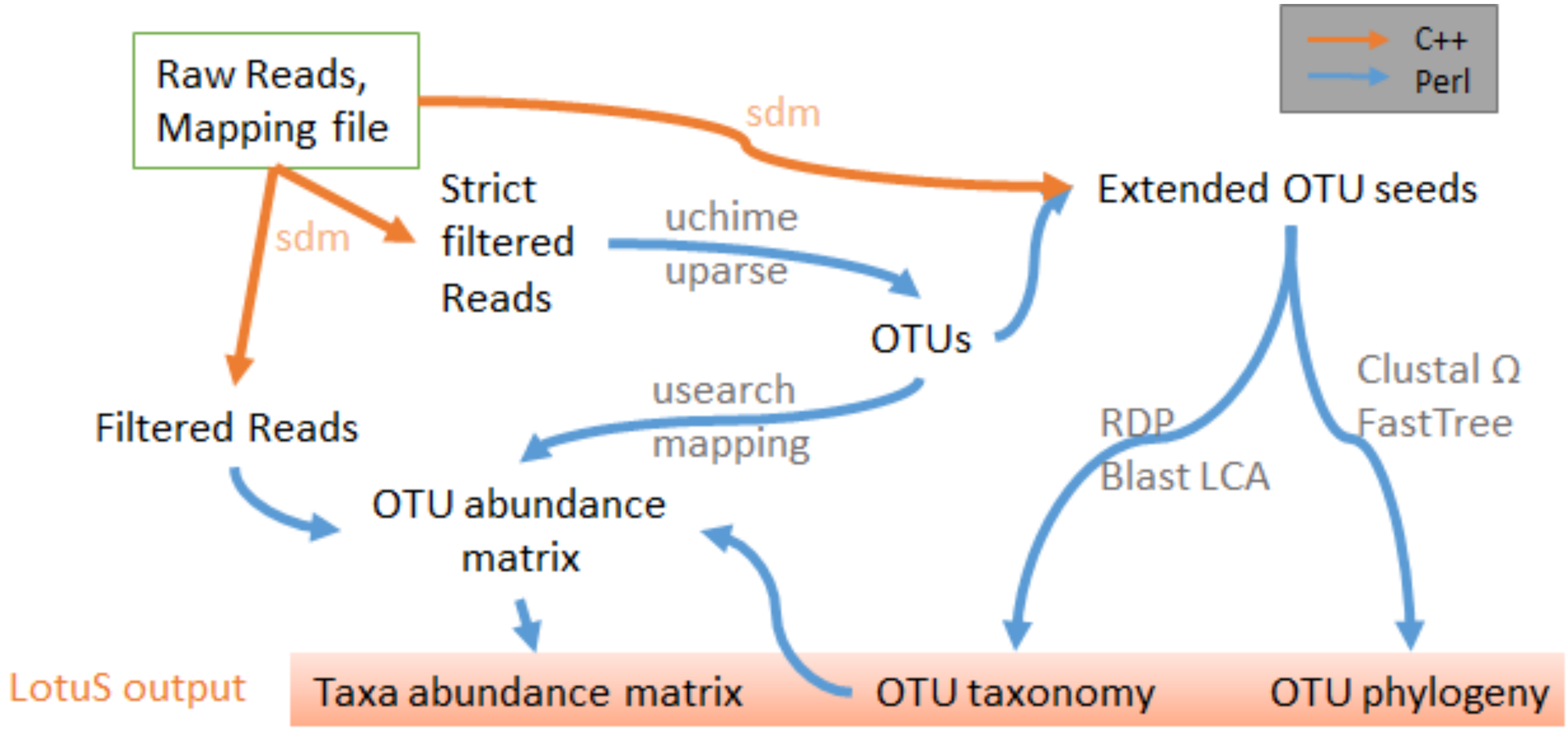
Lotus

See workshop!

- developed by Falk Hildebrand and collaborators
- designed with speed and ease of use in mind
- extensive quality filtering and demultiplexing options provided by **sdm**
- Robust LCA algorithm to assign taxonomy
- Analysis of R1 and seed extension later (quality strategy)
- Simple output folder with OTUs, OTU table and taxonomy table

http://psbweb05.psb.ugent.be/lotus/

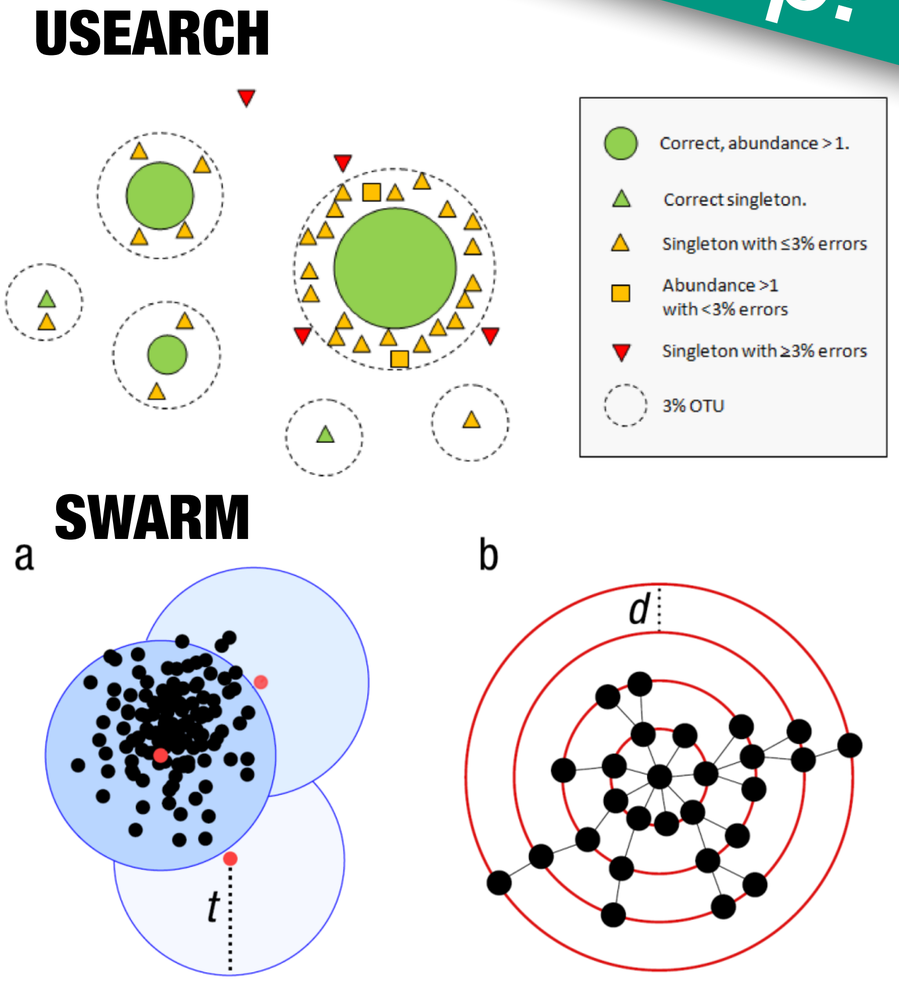
Lotus


C++
Perl
Raw Reads,
Mapping file
sdm
sdm
Strict
filtered
Reads
uchime
uparse
OTUs
Extended OTU seeds
usearch
mapping
Filtered Reads
RDP
Blast LCA
Clustal Ω
FastTree
OTU abundance
matrix
LotuS output
Taxa abundance matrix
OTU taxonomy
OTU phylogeny

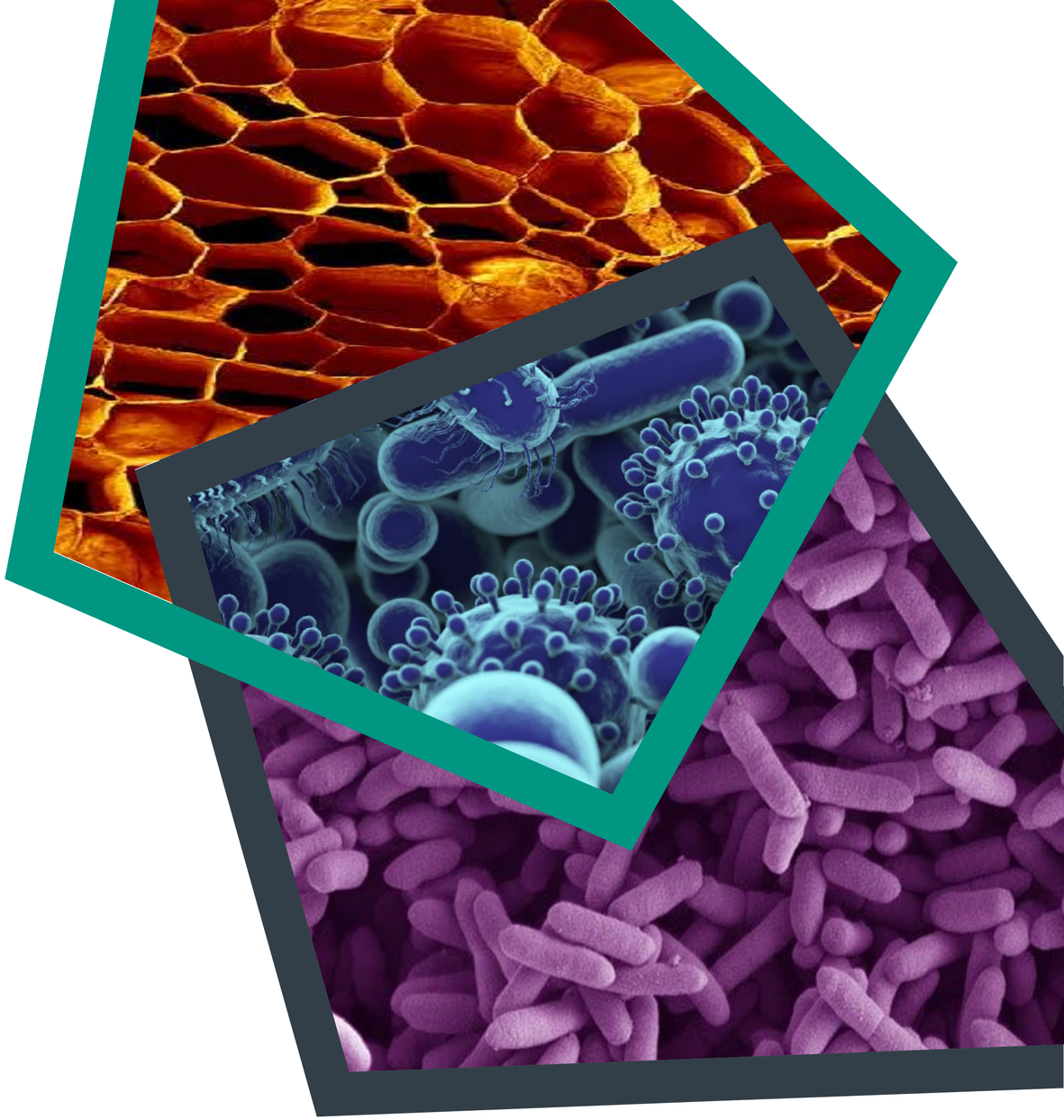
See workshop!

Lotus

- Clustering can be done with **UPARSE**, **CD-HIT** or **SWARM**
- Taxonomy: multiple DBs and algorithms (RDP, BLAST, LAMBDA)
- Ships a fast quality filtering and demultiplexing program (**sdm**): clustering input is a set of clean seqs



Quadram Institute Science Health Food Innovation



Hands on Workshop

https://github.com/telatin/lotus-tutorial

Metadata

A “Qiime” like Mapping File

Input Formats

“sdm” can be used to demultiplex and quality filter

Sequence Analysis

Lotus

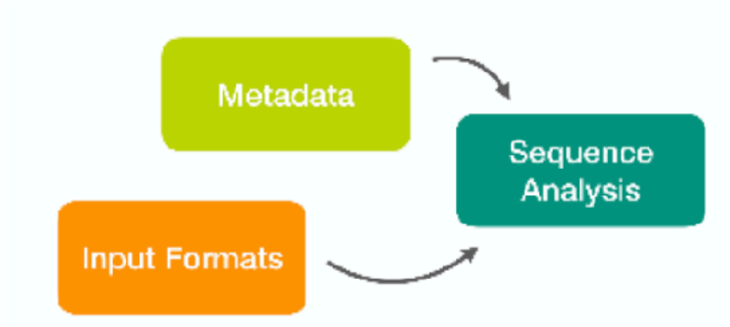
Don't try everything

Lotus tutorial

Lotus is a suite of tools to automatically end efficiently analyze 16S datasets.

This repository contains some documentation and hands out notes (see below), as well as three datasets to pra
(in the datasets subdirectory).

This workshop focuses on a common triad of tasks: dealing with different input formats/technologies, propagati
making use of metadata (mapping files), and performing the 16S analysis itself.



16S analysis with Lotus

👶 **First steps: running lotus**

- Running Lotus

📂 **Preparing input from different sources**

https://github.com/telatin/lotus-tutorial

Metadata

It's a TSV file with a fastq with the **fastqFile** field linking to the input files.

#SampleID	Facility	fastqFile
13xSPFxCD	SPF	13xSPFxCD_R1.fastq,13xSPFxCD_R2.fastq
14xSPFxCD	SPF	14xSPFxCD_R1.fastq,14xSPFxCD_R2.fastq
15xSPFxCD	SPF	15xSPFxCD_R1.fastq,15xSPFxCD_R2.fastq
16xSPFxCD	SPF	16xSPFxCD_R1.fastq,16xSPFxCD_R2.fastq
19xSPFxHFD	SPF	19xSPFxHFD_R1.fastq,19xSPFxHFD_R2.fastq
21xSPFxHFD	SPF	21xSPFxHFD_R1.fastq,21xSPFxHFD_R2.fastq
22xSPFxHFD	SPF	22xSPFxHFD_R1.fastq,22xSPFxHFD_R2.fastq
24xSPFxHFD	SPF	24xSPFxHFD_R1.fastq,24xSPFxHFD_R2.fastq

Metadata

- Lotus provides a script to generate a minimal mapping file (autoMap.pl), but it will only contain the paths and samples will have progressive names (e.g. SMPL1, SMPL2…)
- Its output can be edited (also with Excel) to add the extra columns and then copied and pasted in **text** format with tabs
- Can be checked `lotus.pl -check_map  MAPPING.TXT`

```
#SampleID      Facility        fastqFile
13xSPFxCD      SPF             13xSPFxCD_R1.fastq,13xSPFxCD_R2.fastq
14xSPFxCD      SPF             14xSPFxCD_R1.fastq,14xSPFxCD_R2.fastq
```

Input Formats

- Common scenario: a set of FASTQ files already demultiplexed (Illumina MiSeq, Illumina HiSeq…)
- Demultiplexing: you have an Illumina run with the barcodes (indexes) in a separate file: **sdm can demultiplex**
- Old datasets: you download a .fna and .qual file of a 454 Run, that had the barcode (MID) at the beginning of the sequence: **sdm can demultiplex**

Analysis

lotus

- Basic usage:

```
lotus -i input_folder -m mapping.txt \
    -o output [-c conf.file -s sdm.file]
```

- The **configuration** file allows to set several options, being a “protocol” file. There is a template file in the lotus installation dir
- Since Lotus can handle demultiplexing, you can supply demux and quality filtering parameters with an sdm file

Analysis → lotus

- Full documentation: link to lotus.md

- `-simBasedTaxo [blast, lambda…]` to change classification algorithm (default: RDP)
- `-refDB [GG, SLV…]` to change reference database
  - a custom database can be used (see tutorial)

Analysis

in the Lotus folder

- `lOTUs.cfg`: the configuration file of Lotus, containing the links to external tools and databases
- `sdm_*`: a set of pre-made “sdm” option files, that can be used as templates for the different sequencing platforms
- `sdm`: the binary of “sdm” is also there
- `autoMap.pl`: a tool to generate a “basic” mapping file

Unfortunately…

Quadram Institute Science Health Food Innovation

Thank you!

Andrew Page

Nabil-Fareed Alikhan

Thanh Le Viet

Leonardo De Oliveira Martins

Falk Hildebrand

Clémence Frioux

we are hiring